A Model of Speech Perception by Humans

L. V. Bondarko, N. G. Zagorujko, V. A. Koževnikov, A. P. Molčanov, and L. A. Čistovič

Translated from Russian by Ilse Lehiste*

*Sponsored in part by the National Science Foundation through Grant GN-534.1 from the Office of Science Information Service to the Computer and Information Science Research Center, The Ohio State University.

A Model of Speech Perception by Humans

L. V. Bondarko, N. G. Zagorujko, V. A. Koževnikov, A. P. Kolčanov, and L. A. Čistovič

Translated from Russian by Ilse Lehiste

Foreword

This book sets forth some results of investigations in the areas of psychology, physiology, and experimental phonetics, directed towards the elucidation of the mechanism of speech perception by humans. On the basis of these data and the application of methods of the theories of complex systems and pattern recognition, a plausible model of speech perception by humans is presented.

The work may be of interest to specialists working in the area of the automatic recognition of speech signals: mathematicians, engineers, physiologists, psychologists, and linguists.

1. Introduction

The authors of this work are united in the conviction that the elaboration of a model for speech perception by humans coincides in practice with the elaboration of a system of automatic recognition of a sufficiently large set of speech events.

It is not necessary (and, for the time being, not possible) to demand complete structural isomorphism between the human speech perception system and the system of automatic recognition of speech signals. One can, however, hope for a functional resemblance between the model and the original.

In the process of developing the model, it is unavoidable that questions arise which are inaccessible (or accessible with difficulty) to direct experimental investigation. Insufficient information is then supplemented by guesses and assumptions. The first natural test which these assumptions must meet consists in the requirement that the model which has been set up using these assumptions must be efficient. This, of course, cannot be established before the model is converted into a technical construct or a machine algorithm.

Is it impossible to solve the problem of constructing a model of speech perception in a purely formal way?

For example, let us try to look at the procedure of speech recognition from the point of view of the theory of complex systems [1]. If one does not demand structural isomorphism between natural structures and those to be designed, then it is possible to assume

an infinite number of variants of the automatic speech recognition device. Optimal will be the automaton that will recognize a given lexicon with the required reliability P_o at a minimal cost R. Obviously R will be a function of the cost of memory elements (short-term as well as long-term memory) and other elements which enter the construction.

Hardly any speech researcher believes at the moment that a sufficiently large set of words can be recognized immediately from current parameter values of the speech signal. It is the experience of many laboratories that this method of approach is justified only when the vocabulary does not exceed 20-25 words.

For more complex tasks, multi-stage hierarchical structures of recognition devices are usually proposed. A general block diagram of a multi-stage recognition device is presented on Fig. 1.

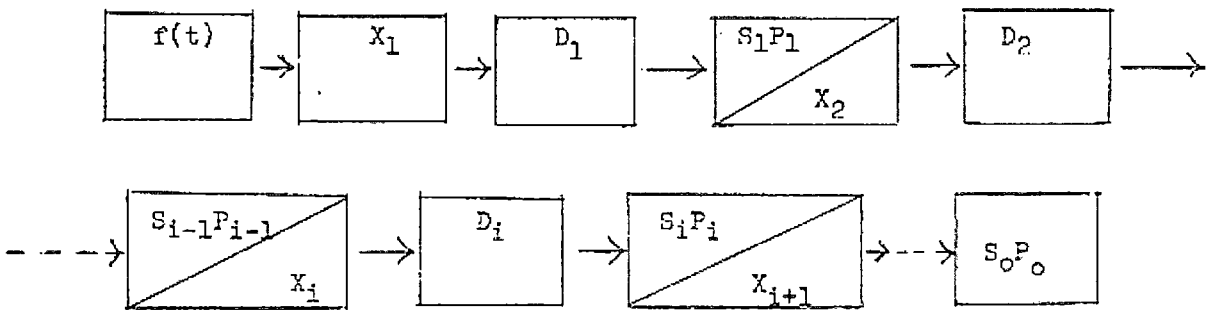

Figure 1

Here X_i constitutes the description of the signal at the input to the i-th level of perception. The classifier D_i, with the help of certain rules ("decision functions") makes the decision whether an unknown vector-realization X_i belongs to this or that element of the alphabet S_i with a reliability P_i. The sequence of elements S_i constitutes the selection space X_{i+1} of the next (i + 1th) level of the recognition device.

The diagram given on Fig. 1 is a general scheme which may serve as a skeleton model for the analysis of existing artificial as well as natural hierarchical recognition devices. The analysis is reduced to determination of the number of levels and the structural elements of each level X, D, S, and P, and the nature of the interaction between these elements.

Externally given requirements for the planner are usually the speech signal f(t) at the input, the lexicon S_o, and the reliability P_o at the output of the system. All intermediate blocks may be selected arbitrarily.

First of all, it is indispensable to implement the transition from f(t) to a more compact and at the same time sufficiently informative description X_1. It is possible to try several, for example "k_1", variants of description. Certain short speech

elements S_1 will be recognized according to this description with a reliability P_1, after them--elements S_2 in their turn, etc. Classifiers may have a different structure at every level. It would be desirable to examine several ("r") versions of algorithms for decision-making.

There are no formal limitations on the intermediate alphabets S_i. At every level it is possible to examine the suitability of "q" variants of the alphabet. If the number of stages equals m, then N different variants of recognition systems are subject to examination, whereby

$$N = k_1 \times r_1 \times q_1 \times r_2 \times q_2 \ldots\ldots r_m = k_1 \times r_m \prod_{i=1}^{m-1} r_i q_i .$$

For the sake of simplicity, let us assume that $k_1=r_i=q_i=n$. Then $N = n^{2m}$, and when $m = 4$, and $n = 5$, $N = 360{,}000$. It is clear that the examination of so many variants is practically impossible, especially considering the fact that each variant represents a very cumbersome task. The most expedient is the following 'biotic' approach: for a first approximation to the optimal schema, the variant should be selected which incorporates all reliably known facts concerning the physiology and psychology of speech perception. Later on it might be attempted to find the best approximation to the optimum in the neighborhood of the point represented by this variant.

As was noted above, the general operating criterion for testing the quality of systems being projected is the summary complexity (cost) of the system R, complying with the limitations $f(t)$, S_o and P_o. It is possible to determine the complexity of each stage separately. For example, the complexity of the classifier D_i is a function of such quantities as the extent of the selection space (i.e. the number of elements in the lexicon of the preceding stage S_{i-1}), the number of elements to be recognized S_i and the type of the decision function D_i. The selection of D_i depends in turn on the required reliability of recognition P_i and on the level of reliability P_{i-1} with which the elements S_{i-1} had been recognized. Therefore

$$R_i = f(S_{i-1}, P_{i-1}, D_i, S_i, P_i).$$

Unfortunately the shape of this function is unknown: it is not known how the reliability of recognition Pi-1 is related to P_i, that is, what mistakes in the recognition of elements S_i are induced by mistakes made in the recognition of elements at the preceding level S_{i-1}. At any rate, the form of this dependence must be determined experimentally (that is, it is necessary to construct and test a concrete automatic machine). This constitutes the basic reason why up to now it has not been possible to optimize a hierarchical structure by formal methods of optimization of the type employed in linear or dynamic programming. For setting up a more detailed schema of the automatic machine at each level it is therefore indispensable to return to known facts about the perception of speech signals.

Thus both the investigators of speech perception (phoneticians, psychologists, physiologists) and the specialists in the automatic recognition of speech (mathematicians and engineers) are now equally interested in setting up a first version of a speech perception model which would incorporate in an unequivocal manner positively known facts about speech perception by humans. The present paper reflects the first stage of our collective efforts in this direction.

In the beginning of the paper (§2) facts and assumptions are presented regarding the structure of a model of speech perception. This is followed by an exposition of some elements of this model (§§3, 4, 5). The paper concludes with a schema and description (§6) of a plausible (from our point of view) version of a model of the recognition of speech signals.

2. Structure of the Speech Perception Model

As soon as we define the final result of speech perception as understanding the meaning of the communication, and demand that it should be possible to understand sentences that have never been heard before, it becomes obvious that the process of perception must be hierarchically organized.

In order to understand the meaning of a sentence it is indispensable to have at one's disposal a description of the syntactic structure of the sentence. In order to carry through a syntactic analysis, it is indispensable to have the sentence first divided into words, and to have assigned to each word its lexical and grammatical characteristics. In order to analyze a word, it is preliminarily necessary to have at one's disposal its phonemic or near-phonemic description. Finally, in order to transform a speech signal into a sequence of phonemes, it is first of all necessary to distinguish in that signal those acoustic features that differentiate phonemes from each other. Distinguishing among acoustic features presupposes an earlier time-frequency analysis of the stimulus.

Specialists engaged in the automatic analysis of texts consider it to be sufficiently well established that the transformation of the alphabetic form of a sentence into a description of its meaning must consist of three successive stages, represented on Fig. 2 to the right of the dashed line [2, 3, 4]. It is obvious that the same three stages must also be present in the analysis of spoken language. Furthermore, the process of the perception of spoken language must include at least two additional preliminary stages of transformation [5, 6, 7, 8].

The first of these stages is the auditory analysis of the speech stimulus. As a result of the operation of this stage, the stimulus is described in terms of acoustic (auditory) features. Logically it is to be expected that the set of features which the auditory system distinguishes in the signal is quite large and is intended for the totality of acoustic signals with which the organism has to deal. It is probable that only a part of these auditory features will turn out to be useful for speech recognition.

The next stage in perception is the phonetic interpretation of the stimulus. The description produced at the output of this block must be already sufficiently abstract and applicable to either an acoustic or an articulatory representation of the speech event. Such an abstract description might be given in terms of, say, phonemes or distinctive features [9].

2.1. The hierarchical model and the possibility of its realization with the help of simple automata.

From the point of view of the theory of complex systems, one of the advantages of hierarchical structure is the fact that each block can be relatively simple ("cheap"), can make do with small amounts of short-term and long-term memory and with a limited number of operations in decision-making.

This is connected with the fact that each preceding block serves as an information filter with respect to the following block, decreasing the dimensions of the signal and bringing it closer to a form that is more convenient for further processing.

Let us imagine that classifier D_i has been allotted a limited number of cells of short-term memory and a limited number of operations. Then it must inevitably be simple, for example linear, and must operate within a restricted space, and the level i - 1 must output such elements S_{i-1} that short sentences thereof can be recognized at the i-th level with the help of linear decision functions. The possibility of replacing a complex decision function with a sequence of simple (linear) classifiers is demonstrated in reference [10]. If the summary complexity (cost) of a multi-stage system with an identical reliability (P_o) in the recognition of elements S_o turns out to be less than the complexity (cost) of a single-stage system, then a re-coding may be considered justified.

To what an extent are these arguments in favor of hierarchical structure supported by facts about the auditory analysis mechanism?

It is considered to be sufficiently well established at the present time that the capacity of the human short-term memory is very limited [11, 12]. This is revealed by data concerning the retention of speech or speech-like stimuli. Thus it has been shown that a sequence consisting of only three vowels or pure tones is remembered as a sequence of decisions about stimuli and not as a sequence of auditory descriptions of stimuli [13]. This makes it possible to believe that the automaton carrying out a phonemic interpretation of the stimuli must perforce work with auditory descriptions of very short segments of the speech train, certainly shorter than the average duration of a word. It is known that the length of a sequence of nonsense words which a human can remember does not exceed 7 - 10 syllables [11, 14]. This obviously characterizes the dimensions of the "temporal window" through which the utterance is "seen" by the automaton performing the morphological analysis of the word. The sequence of grammatically and semantically unconnected words which a human can reproduce after one hearing is likewise very limited [12].

Finally, it has been demonstrated that it is easier to recall the meaning of a sentence than the complete sequence of words constituting the sentence [15]. The adduced data allow one to assume that from the point of view of the total cost of short-term memory at all levels, the hierarchical system of speech recognition must prove sufficiently economical.

Let us proceed further. It is known that the complexity of the classifier depends to a very high degree on the number of patterns to be recognized. Even if we assume that we should succeed in using alphabets of small dimensions at intermediate stages, nevertheless at the last stage the alphabet of objects to be recognized cannot be smaller than, say, the number of words in the lexicon S_o. Would it not be possible to recognize a word without comparing its complete description with every standard item contained in the lexicon S_o?

In reference [16] an algorithm is described of a step-by-step reduction of the lexicon in the recognition process, which is based on the method of "crossing out" proposed by L. Čistovič. In this method, a feature is selected (the first one that occurs or the first in importance among a number of simultaneously occurring features), and all words that lack this feature or this particular meaning of the feature are crossed out from the initial lexicon S_o. Thus the lexicon is sharply reduced. The same operation is performed with other features. The task becomes simpler at every step. At a specified stage, the algorithm proceeds to a comparison of the word with the standard forms of the remaining words in the lexicon, in the complete, multi-dimensional description space, with the help of any chosen decision function. This "combined" algorithm enables one to reduce the decision-making time by several orders of magnitude. There exist reasons to assume that humans follow an analogous procedure. It would be important to find out how concretely it is realized at every hierarchical level of human perception.

An analogous role--reduction of the initial lexicon on the basis of incomplete preliminary information--is probably played by the phenomenon called "psychological setting"--the increase of the a priori probability of certain hypotheses as compared with others. The algorithmic model of this procedure differs hardly at all from the "crossing out" procedure and is possibly realized in living systems by means of a general physiological mechanism.

The reduction of hypotheses and numbers of variants apparently plays an important role at every level of the system, which makes it possible to employ economical classifiers.

A study of the nature of decision functions used by humans has shown that in the process of making a decision in a multi-dimensional space of features, they employ hyperplanes parallel to the planes of coordinates, i.e., the simplest type of linear decision functions [17]. An analogous result was obtained in experiments dealing directly with the perception of speech signals [18]. Consequently, experimental facts support the arguments (the small capacity of short-term memory and the simplicity of classifiers) used to justify the advisability of the hierarchical structure of recognition.

2.2. The hierarchical model and the reliability of recognition.

It is natural that at every level of the hierarchical system information losses must take place, which seemingly makes such a system less effective with respect to the reliability of recognition in comparison with a single-stage system. This question has been repeatedly discussed in the literature in connection with the problem of 'decision units' in speech perception [19]. There exists a large amount of trustworthy experimental data, which demonstrate that in the interpretation of a speech stimulus one relies not only on its acoustic properties, but employs also information concerning phonological and syntactic rules, frequency characteristics of the lexicon, and semantic rules (cf. the survey in ref. 5). It can be concluded from this that one does not make decisions about individual phonemes in the stream of speech and that the units with which one operates correspond to words or even larger segments [19].

If we should interpret this result to mean that acoustic images of whole sentences must exist in human brains, we would arrive at complete absurdity, since we would be forced to assume the presence of images of sentences that have never yet been heard. A reasonable explanation of this result might be the following: if the information at the input to a given classifier proves insufficient, the classifier outputs several possible interpretations of the input signal indicating their a posteriori probabilities, and the final decision in the sense of the selection of one definite alternative may be postponed from stage to stage all the way up to the last one--the recognition of the meaning of the utterance.

The presented ideas correspond to the conclusions drawn by Galunov [20] on the basis of an experimental investigation of the perception of speech in noise. The author arrived at the conclusion that humans carry out a continuous re-coding of the speech stream into phonemes, but make final decision after the elapse of sufficiently long segments.

There is no doubt that the stability of spoken communication among humans in spite of interference is based on the use of redundancy. Voloshin worked out an algorithm for increasing the reliability of recognition at the expense of the redundancy of the signal [21]. Experimental testing of the algorithm showed that it is indeed possible to build reliably functioning devices for the recognition of oral commands on the basis of phonemes that are recognized with a low reliability.

The complexity (or the reliability) of the automatic recognition device depends strongly on the nature of the distribution of the totality of objects to be recognized in the selection space. Usually, given the recognition reliability P_i of elements of the alphabet S_i, the complexity of the classifier D_i increases with a greater dispersion of those elements S_i in the space X_i. An increase in the dispersion of speech signals usually accompanies an increase in the number of speakers who participate in the experiment.

It would be possible to decrease the dispersion, if one could successfully limit oneself to working with standard forms produced by one speaker, adapting them to the particular characteristics of any other speaker. The possibility of such a procedure was examined in [22]. It turned out that the quality of recognition is in fact substantially increased when a standard is provided for the speaker.

The use of a hierarchical recognition system may allow one to make use of the most varied kinds of information about the speaker, starting from the acoustic peculiarities of his pronunciation and ending with the sphere of concepts with which he operates. One possible mechanism could be changing the a priori probabilities of output units in the lexicons of classifiers. It is possible to think that some kind of elementary adaptation to the speaker is already incorporated at the level at which auditory features are isolated. Thus the phoneme boundary in the space constituted by the first two formants of isolated synthetic vowels depends on the frequency of the fundamental tone and on the frequency of the third formant [23]. The reliability of recognition is also obviously increased through adaptation with respect to tempo, speech loudness, acoustic characteristics of the room, etc.

2.3. Special characteristics of the proposed model

It follows from everything said above that a complete model of speech perception must include such higher stages of information processing that are currently being investigated by specialists in machine translation. The realization of such a complete model in the form of automatic algorithms is hardly possible in the near future. At the same time it is obvious that partial models, describing the transformation of information at separate stages of the chain, should preferably be worked out in such a way that they could later be easily inserted into one general model. For that purpose it is indispensable that the output signals of models of lower levels be identical with input signals to models of higher levels.

At the present time specialists in machine translation work with written texts, and input signals for their algorithms are written words, i.e., strings of letters separated by spaces. In oral speech there are usually no pauses between words, and the problem of determining what is a word appears to be sufficiently complex in itself. Besides phonemic information there is also prosodic information which likewise must be transmitted in a transformed shape of some kind to the input of the block that carries out the syntactic analysis. This compels us to assume that the model for morphological analysis (block 3 on Fig. 2) must be worked out specially in conformity with requirements for oral speech, and that this is a task for the joint efforts of specialists in automatic speech recognition and specialists in machine translation.

Correspondingly, we shall formulate the task of the present investigation as producing a model of the chain of transformations that ensure the transition from an acoustic speech signal to its

description in terms of a sequence of words, in which every word in its turn is described in terms of the set of its lexical and grammatical features. This corresponds to the first three blocks on the schema presented on Fig. 2. Besides that, a word must be assigned at the output certain supplementary prosodic characteristics (this question is not at all clear yet and requires special investigation).

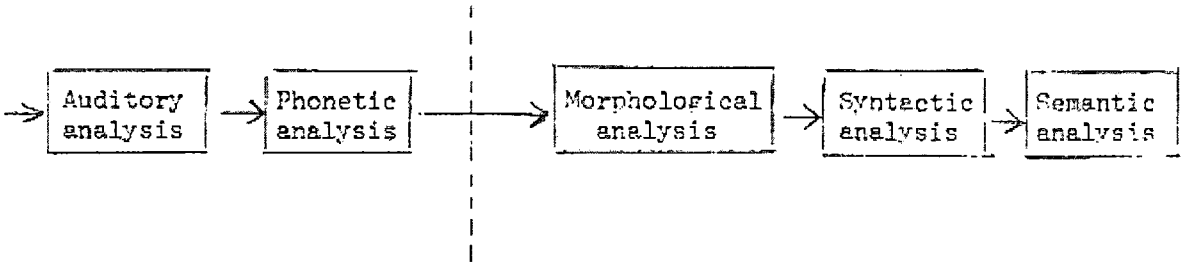

Figure 2.

In the beginning stages the size of the lexicon of the recognition device will rationally be limited to two-three thousand words.

Below we will attempt to formulate some considerations regarding the structure of the first two blocks of the model. Without touching the structure of the third block, we will at this time only define the character of its input signals. It is possible to consider them as sequences consisting of no more than ten syllables each. A syllable will be defined as an element ending in a vowel and containing no more than one vowel. Each syllable is described in terms of phonemes (which, however, may be incompletely recognized). In addition, each syllable is provided with supplementary information characterizing its stress and intonation (the number of degrees and the form of description of these features has not yet been determined).

It has been shown in the work of Lisenko [24] that on the basis of such input information it is, in principle, possible to arrive at a segmentation into words and carry through their morphological analysis.

A specification of the input signals for the third block is indispensable within the framework of this study, since it determines the requirements which must be imposed on the phonetic interpretation block (block 2 on the schema presented on Fig. 2). Let us now turn to facts which are known from electrophysiology of hearing and from psychoacoustics.

3. The processing of speech signals by the auditory organ.

The initial analysis of signals by the auditory organ takes place in the cochlea. Because of the non-uniform structure of the basilar membrane, the transmission of energy from the signal to its various points is realized with dispersion in frequency and time.

In practice, a spectral analysis of the signal takes place in the cochlea according to the transmission functions of the basilar membrane. An approximation of those to Bekesy's data was carried through by Flanagan [25]. An examination of the results of the approximation shows that the representation of the stimulus at a given point of the cochlea is connected with a definite frequency selection and temporal delay of the signal.

It is known from psychoacoustic experiments that a decision about the stimulus is taken with regard to its development beyond the so-called critical time of analysis, consisting of approximately 150-200 milliseconds. Taking into account what has been said, the representation of the signal at the level of the cochlea may be described by means of a contour in the frequency-time-energy space. A given contour reflects the stimulus which immediately brings about the excitation of neuron endings found in the organ of Corti. However, the transmission of this excitation to successive sections of the neural net proceeds differently for different elements of the contour causing the excitation. Psychoacoustic experiments demonstrate that a human listener identifies sound signals as similar if they differ only in amplitude. Invariance with regard to amplitude (loudness normalization) is evidently connected with the transmission function involved in the transmission of the excitation into the neural net. The frequency of impulsation (or the probability that a response occurs after a given time segment) of a peripheral neuron is connected with the intensity of the stimulus that acts upon the corresponding receptor with the logarithmic dependency relation

$$p(x) = \log \frac{1}{t} \int_0^t f_t^2(t')\, dt' \sim \log (E + P_o); \qquad (1)$$

where $p(x)$ - the probability that an impulse will occur in response to the stimulus,
P_o - the quantity accounting for the spontaneous activity of the neuron,
t - integration time,
$f_t(t)$ - the size of the instantaneous value of the stimulus (only its positive part is taken into account).

If the part standing under the logarithm sign in equation (1) is taken to be connected with the average energy of the stimulus, then the response reaction of the neuron at the periphery of the auditory system will be proportional to the logarithm of the average energy of the stimulus E for the time t. The working rule of the elementary structure that isolates any given feature which is dependent upon the spatial distribution of the energy of the stimulus along the basilar membrane and independent of changes in intensity, may be formulated in the following way:

$$A_i = p(x_1) - p(x_2) = \log \left(\frac{E_1(x)}{E_2(x)} + E_o \right) , \qquad (2)$$

where x_1 and x_2 - two arbitrary points on the basilar membrane,
E_o - a constant taking into account spontaneous activity.

It is possible to write an analogous expression for describing the amplitude changes of the stimulus with respect to time. A complete description of the shape of the stimulus, invariant with regard to its intensity, may be obtained if the values represented in equation (2) are formulated for all distinctive points of the excitation contour. It seems that the just presented normalization process of the description of the signal with respect to intensity takes place at the outermost periphery of the neural net and constitutes a part of the mechanisms for isolating the most distinctive points. It is of great interest to examine electrophysiological data as to which elements of the signal are observed to produce the most clear-cut reactions of the neurons in the various sections of the neural part of the organ of hearing.

At the present time, abundant data are available regarding the responses of single neurons, starting from bipolar cells and ending with the cortical part of hearing. It is difficult to say to what an extent the characteristics of these responses are connected with the results of psychoacoustic experiments. Nevertheless, the fact that there exist selective responses of neurons to signals of a particular form constitutes evidence that at least at the periphery of the system of hearing, a representation of the signal is formed which is based on isolating its specific features.

In reference [28] it is demonstrated that there exist two groups of neurons in the cochlear nucleus which can be clearly separated according to the nature of their responses, conditionally called tonal and phasal. The former are characterized by a substantial sharpening of the frequency-threshold curves when the duration of tonal emission is increased from 2 to 100 msec, by a significant dependence of the latency period on the intensity of the signal, by the preservation of the response during the whole length of the stimulus, and also by a clearly expressed temporal summation of the energy of the stimulus. The character of the frequency-threshold curves of the neurons with tonal response itself points to a significantly sharper reaction of the observed neuron to the stimulus at a given frequency than could have been expected on the basis of the frequency characteristics of the inner ear's mechanical system (the so-called sharpening effect). Among the existing hypotheses set up to explain the sharpening effect, that one appears best founded that proposes the existence of some kind of lateral inhibition at the periphery of the neural part of the organ of hearing.

The simplest variant of lateral inhibition, conditioning the sharpening effect, appears to be a scheme for isolating the differences in the intensity of excitation of adjacent elements. For carrying through this operation, the existence of neuronal structure is

postulated that reacts to non-uniformity in the distribution of energy in the space of the receptors.

In the limiting case, such a structure could consist of a single element of a neuron, if the response to a stimulus is proportional to the sum of the absolute values of the differences in the influence exerted on its dendrite system by the receptors. If this is so, then the response of the neuron will be proportional, in the limiting case, to the derivative of one or another order of the function that describes the spatial distribution of energy in the studied section of the receptive field.

An examination of the model of such a scheme of excitation allows one to note the following of its properties. The quality of frequency-selective characteristics, extracted when a tonal signal is fed in at the input of the model, can raise by an order of magnitude the quality of analogous characteristics of input filters (in this case, the frequency characteristics of the basilar membrane). The angle of the slope of the frequency-selective characteristics may reach a magnitude of the order of hundreds of decibels per octave [26]. If two tonal signals act upon the input of the model simultaneously, of which the second is out of tune relative to the mean frequency, one observes a clearly expressed masking of the first signal. The response of the model is insignificant when a signal with a continuous uniform spectrum is fed in at the input. When tonal signals with varying duration are used as stimuli, the model displays clearly expressed effects of temporal summation: an increase in the duration of the signal is accompanied by a lowering of the threshold of exhaustion and an increase in the quality of the frequency-selective characteristics. The latency period of the response depends strongly on the intensity of the stimulus.

The quoted data show that the model of lateral inhibition in the given formulation possesses the basic properties of neurons with tonal response. The characteristic property of the described lateral inhibition model is the sharp isolation of extremes in the spatial energy distribution of the signal. This allows one to propose that the mechanism for isolating formants in speech signals as starting-point features operates with data about the position of extremes in the continuous spectrum of speech elements.

The neurons which give a so-called phasal response to stimuli are characterized by the independence of the response latency period of the signal intensity, by a small dependence of the response threshold and the sharpness of frequency-threshold curves on the duration of the stimulus, and by a significant dependence of the response on the steepness of the onset of the signal. The phasal response usually consists of one or a few bursts, immediately following the onset of the stimulus. Neurons with phasal response constitute about 20% of all investigated elements in the cochlear nucleus. Besides that, their procentual share increases in the higher sections of the neural net of the auditory system (in the inferior collicula etc.).

The described properties of neural structures with phasal response make it possible to assume that they play the role of determining the moments at which a change takes place in the energy

of the signal, concentrated in one or another frequency region. One may assume that a temporal segmentation of the uninterrupted stream of speech is worked out in higher sections of the auditory system on the basis of signals which have been received from phasal-type neurons. Apparently there exist in the neural net of the auditory system isolators of significantly more complex characteristics of stimuli, which describe in detail how their energy changes with respect to time as well as to frequency. The same isolators accomplish the quantitative evaluation of the shape of the perceived signal.

In reference [28] it is shown that in the inferior collicula of rats there are neurons which respond with a group of impulses to a short signal (of the order of 1 msec) and do not respond at all to longer-lasting stimuli (longer than 10 msec).

In the auditory part of the cortex neurons have been isolated that respond only to stimuli whose frequency changes in one or the other direction [30].

The neurons of a given group can be differentiated into three groups according to the nature of their response to a frequency-modulated signal.

The first group is constituted by elements that react to a change in the frequency of a tone in the direction of a higher frequency.

The neurons of the second group respond only when the frequency of a tonal stimulus is decreased (with a definite speed!). Finally, neurons of the third group react to a change in the frequency of a tone, if the direction of this change leads to a closer approximation of the frequency of the tonal signal to the characteristic frequency of the given neuron. A multitude of data, accumulated in investigations of the visual analyzer, point to the existence of structures in its neural part that isolate from external stimuli such elements like contours, angles and more complex configurations [27]. A special significance have those cases in which neurons in the cortex respond only to a spatial movement of a certain kind of signal. It is interesting that the underlining of contours in drawings is realized at the periphery of the visual analyzer of a frog owing to the interaction of the processes of excitation and inhibition, which evolve in time according to different laws [27]. This points to the subtlety and complexity of the structures which accomplish the isolation of informative elements in images. In view of the proposed communality of the basic principles of the processing of information in different sensory systems of the organism, it is useful to consider the possibility that there may exist neural structures in the system of hearing that react to the same kinds of features of signals as are reacted to in the visual analyzer.

Finally some hypotheses should be mentioned that propose the existence of neural structures in the auditory system that isolate certain sequences of occurrences of maxima or sharp decreases of energy in the stimulus at different frequencies and at different moments in time. The existence of such structures

might explain the selective reaction of living creatures to certain sounds with a complex spectral and temporal structure. According to current views, neural structures that react selectively to complex signals are continuously being formed in the neural net in the process of developing conditioned reflexes. However, it is necessary to distinguish structures that come into being as a result of training in the perception of new signals from those that are genetically consolidated.

It is proposed at the present time that neural structures which appear in the process of learning are to a certain degree connected with the mechanism of memorizing, and that they participate as a part of this mechanism in the decision-making process regarding the recognition of signals in the extreme stages of analysis. On the other hand, inherited neural structures participate in the beginning stages of the perception of stimuli; they isolate, i.e., react more sharply to those elements that are most characteristic for the whole broad class of signals which is taken in by living creatures of a given species in the processes of vital activity.

The proposition that a number of neural structures that isolate complex elements of stimuli are transmitted genetically was proven for the visual analyzer by direct physiological experiments.

It may be concluded from what has been said above that the representation of an external stimulus in the reaction of an ensemble of neurons that constitute the neural net may be considered passive only at the lowest level (the level of cochlear receptors). Later on, a reaction to the stimulus is formed from responses of those neural structures that isolate certain elements of the signal from its complete description. The presence of a given element in the perceived signal is indicated by the response of a neuron or a group of neurons at the terminal point of a specialized neural structure. Thus it is assumed that the isolation of certain physical features of the signals is accomplished already at the periphery of the neural net. The admission of such features is determined by the availability of corresponding neural structures. The latter are evidently developed in the process of evolution.

This reduced representation of the stimulus is transmitted to the succeeding level of the neural system, where, on the basis of identified features, further operations are carried through in the classification and recognition of stimuli. It may be assumed that not all detected features are used at the next level of analysis, but only some of them (more precisely--the minimal number needed for making a classificatory decision with the necessary degree of reliability). The complete set of features is necessary only for solving the most difficult tasks of classification, when the ensemble of stimuli to be recognized is sufficiently large. For ordinary tasks, the use of only some of the isolated features appears to be sufficient. Therefore it is natural to suppose that the selection of necessary features takes place during the process of solving the problem of classification at a relatively higher level of analysis. Consequently there must exist the possibility of transmitting various combinations of features up to the moment at which a reliable answer has been achieved.

A certain decrease in the quantity of transmitted features occurs due to the masking phenomenon. Since the observed processes, from the isolation of features of signals to the making of decisions, are accomplished by means of essentially non-linear transformations of the values characterizing the stimuli, such events take place as the suppression of a weak signal by a strong one when the two act simultaneously, etc.

When we study the spectra of speech sounds (especially those isolated from running speech), we easily find that they contain a large quantity of maxima, minima, sharp decreases, etc. As is demonstrated by psychoacoustic experiments, not all of the isolated maxima have always the same significance for classification. Furthermore, in the process of being isolated many of them are suppressed by more powerful neighbors. Up to now there is insufficient information for establishing the rules according to which certain and not other features (e.g. the spatial distribution of the remaining maxima) are selected by the nervous system in the solution of a concrete classification task. However, the possibility itself for the existence of an operation of surveying various combinations of features appears very probable.

4. The auditory description of the speech signal

The question about the form of the auditory description of the speech signal, i.e., about the set of features and their nature, on the basis of which the signal is characterized in the process of perception, has been left practically unexplored up to now by psychoacousticians and linguists. Almost all information concerning the acoustic properties of speech sounds and their perception has been obtained either by the analysis of differences in dynamic spectrograms of natural speech signals having a different phonemic value, or by means of investigations of the perception of synthetic speech-like stimuli. The method according to which an investigator describes the signals is in both instances predetermined by the properties of the instruments with which he is working. As a result, the terminology used for the description of the features of the speech signals has turned out very specialized and at the same time poorly formalized. This refers to such basic concepts as formant, transition, locus, burst, etc.

The qualitative peculiarities, on the basis of which one may distinguish different speech sounds from each other on spectrograms, are more or less known at the present time. Thus, in order to determine which consonant starts a simple CV-syllable, it is useful to examine the following features of the signal [31, 32, 33, 71]:

1. Presence or absence of the fundamental frequency from the very beginning of the syllable;
2. Abrupt rise in the frequency of the fundamental at the transition from the consonant to the vowel;
3. Presence or absence of noise from the very beginning of the syllable;

4. The frequency position of the spectral maximum of the noise segment or the burst;
5. The bandwidth of the noise:
6. The slope of the increase in intensity of the noise:
7. The duration of the noise segment:
8. The intensity of the noise;
9. The presence of formant structure from the very beginning of the syllable;
10. The lack of discontinuities in formant structure during the extent of the syllable.

Besides that, the so-called formant transitions are of essential value in determining the point of articulation of the consonant and for determining whether it is 'soft' or 'hard' (in Russian).

Even though the listed features have not been formalized (it is not indicated by which methods the features may be detected, nor have decision-making rules been given), establishing the list itself constitutes an important stage in the investigation of speech--the stage of acquiring primary acquaintance with the acoustic peculiarities of the speech signal.

If such a qualitative and, consequently, not very definite description of the speech signal is satisfactory for a number of phonetic tasks and is fairly popular among phoneticians, engineers prefer to use a completely formal, but very inconvenient 'complete' description of the signal in frequency and time. In this process the dimensions of the space in which phonemic decisions are made turn out to be very large, and the decision functions are complex. In addition, such basic difficulties appear as the inevitability of a very precise preliminary segmentation of the speech stream and the indispensable necessity of time normalization.

The hypothesis in favor of which data will be presented below consists of the following. We propose that the complete description of the signal takes place in the cochlea alone; in the processing and transmission of the signal in the neural net, a gradual reduction of information takes place. At a certain level, a transition takes place from the description of the whole envelope curve of the spectrum at a given moment in time (the curve describing the distribution of impulsation along the projection of the cochlea) to the description of the position on the frequency axis (the projection of the cochlea) of a few points (maxima or turning-points) on this curve. On a still higher level, a determination of the parameters of the curves takes place, which describe the displacement in time of the points isolated along the frequency axis. Such assumed parameters might be the direction of movement, speed, magnitude at the turning-point, etc. It is assumed that these already fairly complex characteristics of the signal constitute the features with which the block operates that carries through the phonemic interpretation of the stimulus. It seems to us also that this point of view is not original at all, and that the majority of speech researchers, not discussing it specifically, nevertheless proceed from this standpoint.

For convenience in presenting the material we divide it into three groups of data, concerning the perception of: 1) the envelope

curve of a stationary complex signal, 2) the fine temporal structure of a stationary complex signal, 3) changes in time of the spectrum, fine temporal structure, and the intensity of the signal.

4.1. The perception of the envelope curve of a complex stationary signal

As a rule, speech researchers use the assumption that in perception, humans somehow determine the formant frequencies of the speech signal and employ this information in making phonemic decisions. Correspondingly, stimuli are described in terms of formants, and the results of experiments are presented from the point of view of the dependence of the response probabilities (identification or differentiation) upon the parameters of the stimulus. Since to the best of our knowledge none of the researchers has observed any contradictions in the obtained results, the description of the stimulus in terms of formant frequencies appears adequate enough.

However, a change in the formant frequencies of a synthetic speech-like stimulus signifies a change in the spectral envelope of this stimulus. Therefore the assumption is not excluded that in fact humans employ in perception not the values of formant frequencies, but either the whole spectral envelope (the outline of the distribution of impulsation along the projection of the cochlea) or, for example, the relative amounts of energy in some fixed bands. The latter point of view was proposed by Varšavskij in the discussion of a possible model of speech perception [34].

In order to prove that a human listener in fact measures formant frequencies, it is first of all indispensable to decide upon a sufficiently formal definition of a formant. Further, it is necessary to find some kind of a reaction which would regularly change when a formant frequency is changed, and would not depend on other parameters of the stimulus.

The term 'formant' is used by speech researchers in two different meanings: a formant is understood to be either a pole in the transfer function of the vocal tract, or a maximum in the spectrum of the analyzed sound. In the latter case, what one really has in mind is a maximum on the curve describing the response of the analyzing apparatus to a given sound. It is understandable that such a definition of a formant is already very indefinite; it seems to depend on the properties of the analyzing device.

If the isolation of formants takes place at a relatively low level in the processing of auditory information, it is difficult to assume that complex procedures are employed in the process-- procedures making use of data concerning the transfer function of the vocal tract (e.g., procedures of the 'analysis-by-synthesis' type, cf. [35]). Then the first of the two above-quoted definitions of the formant does not apply, and the formant may, as a first approximation, be identified with a spectral maximum or, more truthfully, with a maximum on the curve describing the response to the signal in those initial links of the auditory system which

perform the spectral analysis.

In psychoacoustics it is accepted that the response curve is sufficiently well reflected in the curve describing the masking called forth by the stimulus (for discussion, cf. [36]). Questions as to whether the auditory system employs the frequency of a spectral maximum of a complex stimulus as a parameter of that stimulus, and in which way it measures the frequency of the maximum, relate to the general problems of the physiology of hearing and of psychoacoustics and are closely connected with questions concerning the mechanism of auditory determination of frequency (hypotheses concerning these mechanisms are summarized in ref. [36]).

In the works of Šupljakov [37, 38, 39] direct proof was obtained that a human listener determines the value of the frequency of the first spectral maximum in natural and synthetic sibilant (fricative) consonants of the type /s/ and /ʃ/. In the given case this maximum corresponds to the second formant (in the sense of pole). It appeared that the frequency of the maximum carries two kinds of information: on the one hand, it determines the musical pitch of the sound and on the other hand, the phonetic category ('hard' or 'soft' consonant) to which a given sound belongs. The connection between the frequency of the maximum and the pitch of the sound may be considered immediate; the decision regarding the 'hardness' or 'softness' of the sound is determined by whether the frequency of the spectral maximum is higher or lower than a fixed threshold. The following constitutes proof that in this case it is the frequency of the spectral maximum that is determined and not some other parameter of the spectral function: a change in the amplitude of the maximum has no significance as long as it is above the detection threshold, and the value of the boundary between 'hard' and 'soft' consonants on the basis of the maximum is the same for /s/ and /ʃ/, which are significantly different from each other on the basis of other peculiarities of their spectra. The precision of the determination of the position of the boundary by subjects employing the method of active search [39] appears very high (1.5 - 3.0%), which suggests that the procedure for auditory isolation of the maximum and for determining its frequency is sufficiently effective.

Data obtained in the investigation of the perception of synthetic whispered vowels [40] indicate that the frequency of a spectral maximum corresponding to the first formant of a vowel is also determined. It was discovered that the boundary between the vowels [i] - [o] and [u] - [ø] in the F_1-F_2 plane is represented by a straight line, parallel to the axis of the second formant. This means that for separating vowels according to these categories, what is employed is the frequency of the spectral maximum corresponding to the first formant, and not the whole spectrum envelope or, for example, the ratio of energy in some fixed frequency bands.

The data quoted above were obtained for the case in which the concept of spectral maximum is not in doubt (the maximum is sufficiently sharp), and the frequency of the spectral maximum coincides with formant frequency defined as a pole in the transfer function of the vocal tract. The question naturally arises what is

being employed as parameters of the spectral function, if the range of isolated frequencies (i.e. frequencies containing the essential part of the energy) is sufficiently wide and does not possess a clearly defined maximum. Some data for replying to this question have been obtained in experiments dealing with the estimation of the pitch of band-limited noise. It has been found that for noise whose spectrum is limited only from one end (high-frequency or low-frequency noise), the perceived pitch is determined by the frequency of the cutoff [38, 41]. For band-passed and relatively narrow bands of noise, pitch is determined by the average geometric frequency of the noise [42, 43]. It is obvious that a further special investigation is needed concerning the width of the band at which the transition from one to the other mode of pitch estimation takes place.

It is a very complex and as yet unexplored question as to which parameters are used for describing the response of the auditory spectral analyzer to a signal with a discrete spectrum (stationary vowels). On the basis of available psychoacoustic data [44] one must expect that when the fundamental frequency of the voice is very low a separate maximum on the response curve of the analyzer must appear corresponding to almost each harmonic in the frequency space below 1000 Hz. There is not one of these maxima that may not coincide with a formant frequency in the sense of a pole in the transfer function. Ths supposition that a human listener perceives the frequency of the strongest harmonic as the formant frequency does not agree with data about the great precision in distinguishing the frequency of a formant (as a pole). According to Flanagan's data [45], the just noticeable difference in the frequency of the first formant amounts to 3%. Thus this question remains unclear at the moment and urgently demands further investigation.

4.2. The perception of the fine temporal structure of a stationary complex signal

In the perception of speech, the decision regarding the character of the source of excitation (voiced, noise-like, mixed, impulse-like) is made on the basis of the fine temporal structure of the signal. Discrimination between consonants which differ among themselves on the basis of the source of excitation is practically not disturbed at all under significant spectral distortions of the speech signal [46, 47]. The question is still open as to which parameters are employed in the auditory system to describe the temporal structure of sound. It is assumed that they must be some parameters of the distribution of intervals among nerve impulses.

When speech sounds are to be classified according to their temporal structure, it is convenient to divide them first of all into two groups on the basis of whether their structure can or cannot be auditorily determined. Existing data allow one to believe that if the duration of the first of a sequence of two adjacent stimuli is shorter than 15-20 msec, a listener does not recognize its temporal structure. When either fricatives (like s) or periodic consonants (like m) are shortened up to that duration, a listener

perceives both as plosives (p or t), which are characterized in speech production by impulse-like excitation [48, 49].

Within the class of stimuli with a temporal structure that can be auditorily distinguished, three categories may be established: periodic signals (with a harmonic spectrum), amplitude-modulated noises, and continuous noises. The amplitude modulation of noise in voiced fricative consonants approaches right-angled modulation (for a discussion of noise production mechanisms, cf. [50]); its frequency is fairly low and corresponds to the frequency of vocal-fold vibration. Psychoacoustic experiments (cf. the surveys in [36, 51]) demonstrate that at these modulation frequencies listeners not only detect them, but distinguish one modulation frequency from another and, furthermore, assign a pitch to the signal that is equivalent to the frequency of modulation. It has been also shown by numerous experiments (cf. the survey in [36]) that the period of repetition of a complex periodic vibration is perceived and serves as the basis for estimating the frequency of the sound, even if the spectrum of the sound does not contain corresponding low-frequency components.

One may thus assume that for the description of the fine temporal structure two parameters are necessary and probably sufficient. One of them must reflect some kind of measure of the degree (explicitness) of periodicity, the other must reflect the magnitude of the period of repetition.

In psychoacoustics, perception of the temporal structure of the signal is sometimes taken to include the perception of changes in the signal that occur with a frequency below that of the vibration of the vocal folds (below 100 Hz). Periodic low-frequency changes in the signal are almost never encountered in speech (the exception is provided by the sound r); however, single occurrences of rapid changes in the spectrum (formant frequencies), fundamental frequency, or intensity level appear regularly. It seems to us that the parameters by means of which these changes are described can be considered as derived from the parameters that have been examined in this section (formant frequencies, period, degree of periodicity). The next section will be devoted to their consideration. The intensity of vocal fold vibration should obviously also be assigned to parameters of the same level as formant frequencies and the period of repetition. The time constant for the auditory determination of intensity is approximately 10-20 msec [52, 53].

The essential characteristic of the parameters listed above is that auditory measuring devices responsible for their detection must have a low inertia. Therefore it is possible to consider these parameters as indicators of instantaneous (current) properties of the speech signal. Only stationary sounds (synthetic or artificially pronounced isolated vowels and some consonants) can be described with a single value for each parameter. In natural connected speech, the values of the parameters of the signal constitute functions that change with respect to time.

4.3. The perception of changes in time of the spectrum, fine temporal structure and intensity of the signal.

An essential characteristic of natural connected speech is the fact that the values of formant frequencies, fundamental frequency and intensity change significantly in the course of the duration of separate speech elements, and that these changes are by no means random, but have a completely regular character. From the point of view of speech production, the stream of speech may be considered as a sequence of open syllables [14, 54, 55, 56]. A standard kind of time function corresponds to each syllable and each parameter (formant frequencies etc.), as well as a limited set of possible transformations of the function, connected with the tempo of pronunciation, intonation, and position of the syllable within a word.

This means that the curve describing the change of each of the parameters in the course of a long utterance may, as a first approximation, be viewed as a sequence of sections (pieces), where each section is a certain standard time function corresponding to a syllable.

The question is almost unexplored as to how a signal is described in the process of perception whose parameters change in time. In order to discuss the small amount of fragmentary data that are available it would be useful first to summarize existing hypotheses.

One of the hypotheses says that the time picture is described completely, i.e. that readings are used for each of the parameters taken at, for example, every 10 msec. Thus the change in the parameter during the extent of the syllable is described by a set of numbers reflecting the value of the parameter at successive discrete instances in time. The difficulties connected with this hypothesis consist first of the fact that such a description appears extremely unwieldy (a large memory capacity is required for registering it) and, secondly, of the fact that phonetic and prosodic information has not yet been separated.

When this form of description is employed, the same syllable produced by the same speaker will look different, depending on the tempo with which it was pronounced, on the position of the syllable within a word, or whether it carries logical stress, etc. Therefore it is still difficult to use such a description as the immediate input to the block that performs the phonemic interpretation.

Two ways have been proposed to overcome these difficulties. In one of these [57], the representation of the syllable obtained at perception that has been entered in short-term memory is subjected to certain (for example topological) transformations, as a result of which it is given a more standardized shape. The formalized representation enters at the input of the block performing the recognition.

The second way [58] consists of comparing the representation entered in short-term memory with a standard syllable which is synthesized under the assumption of a different tempo, position within the word, etc. The phonemic composition of the syllable, tempo, position within the word etc., constitute initial variables for the synthesizer. What is

being sought is such values of those variables with which the synthesized representation and the representation entered into short-term memory are closest to each other.

Although the presented hypothesis of recognition by syllables appears sufficiently logical from the point of view of existing knowledge about the process of speech production, there is no proof whatsoever that perception of speech by humans is indeed accomplished in this manner. Furthermore, there exist data that contradict this hypothesis. These data concern the possibility of partial recognition of the syllable, recognition of some of its phonemic (distinctive) features while others are not recognized (for example, the recognition of the manner of articulation of a consonant without recognizing its point of articulation, etc.), and recognition of prosodic features without the recognition of phonemic ones.

Into this category falls also the fact that some phonemic features can be recognized earlier than others, even before the listener hears the complete syllable [14]. These data suggest that the auditory description of the syllable that enters at the input of the block performing the phonemic interpretation is already organized in such a manner that it allows parallel, multi-channel processing.

The second hypothesis consists of the proposition that the curves which reflect the change of formants, fundamental frequency, etc., during the extent of the syllable are described in perception by means of the set of features of those curves. These features might include the direction of change of the parameter, the rapidity of change, and the value of the parameter at a certain specific point.

In experiments in the perception of synthetic stimuli it has been shown that the nature of the initial transitions of the second and third formants of vowels carries information about the point of articulation of the consonant [60, 61, 62, 63]. At first it was proposed that the characteristic of the transition used by human listeners is its 'locus'--the initial value of the formant frequency, which supposedly does not depend on the vowel with which the given consonant is connected in the CV syllable. (Later it was discovered that the locus value is different for different vowels [64]).

Not long ago Stevens [65] examined in detail spectral changes in the sound during the transition from stop consonants to vowels at different points of articulation of the consonant. Comparing his results with data concerning the perception of formant transitions obtained at Haskins Laboratories, he advanced the hypothesis that when formants are close to each other in frequency, what is determined in perception is the frequency position of the sum of the spectral maxima. From this point of view, the transition from labial consonants to vowels is always characterized by a rise in the frequency of the spectral maximum, and the transition from apical consonants to vowels is characterized by a lowering of the frequency of the maximum. As the absolute frequency position of the maxima depends on the nature of the vowels, it is natural to admit that the useful feature which distinguishes between labial and apical consonants is the direction of the change in time of the frequency of the maximum.

Data in direct support of the assumption that human listeners use the direction of change in the frequency dimension of the energy

maximum as the indicator of the point of articulation of the consonant were obtained in experiments for determining the boundary between p and t in synthetic syllables constructed in the following manner: the syllable consisted of a short (10 or 15 msec) emission of noise with a narrow bandwidth or of a sinusoid, followed by a harmonic signal with a maximum in the range of 300-600 Hz (u) or 900-1200 Hz (a). The subject changed the frequency of the tone (the center frequency of the band-passed noise) in the short emission, trying to find the value at which perception changed from tu to pu (ta to pa) or from pu to tu (pa to ta). It turned out that the boundary between pu and tu is located near 400 Hz, and the boundary between pa and ta around 1000 Hz [66].

The notion that it is the direction of the change that is perceived and not the locus, i.e., the initial value of the formant frequency, is supported by the categorical nature of perception [67]. Subjects behave as if they detected only the presence or absence of a change in the formant and its positive or negative sign.

The transition from a velar consonant (k, g) to a vowel is characterized by the fact that the frequency of the maximum does not change its position in time, but the pertinent frequency range at the beginning of the transition is narrow (the frequencies of F_2 and F_3 coincide) and then becomes wider (the formants separate during the vocalic segment) [68]. In Russian, one of the features distinguishing k - g from t - d and p - b is the greater duration (for k - g) of the noise of the explosion.

There exist also data supporting the notion that humans use the rate of change of formant frequency as a useful phonetic feature [69].

If the direction and the rate of change of formant frequency play the role of useful features of consonants, then, it is proposed, useful features of vowels consist of the values of formant frequencies during the stationary part of the vowel (if present) or at the turning-point of the formant curve.

Of great interest are here experiments in estimating the frequency of a short stimulus (20-50 msec) whose frequency was changed significantly (raised or lowered) during its extent [69, 70]. These experiments showed that humans equate the pitch of such signals with the pitch of a stationary tone having a frequency equal or close to the terminal frequency of the signal. These data permit one to exclude the assumption that it is the time average of the frequency of the changing signal that serves as an auditory parameter.

These experiments should be followed by an investigation of the perception of the pitch of sounds with more complex changes of frequency in time.

Fairly little is known regarding the auditory features employed in describing the curve that represents the change of the fundamental frequency of the voice during the extent of the syllable.

It has been shown that a sudden jump in the frequency of the fundamental at the transition from consonant to vowel serves as a useful feature in distinguishing b from m [71]. The threshold value for the rise in fundamental frequency that corresponds to the boundary between these consonants consists of 10% of the absolute value of the fundamental frequency.

On the basis of a fair amount of phonetic data it may be assumed that a more gradual rise (proceeding with lesser rapidity) in the fundamental frequency during the syllable serves as a feature of stress.

Data are available showing that the relative intensity of a consonant constitutes a useful feature [72] for distinguishing fricatives from each other [71]. However, the question has not yet been investigated at all which and how many parameters are used to describe auditorily a signal changing with respect to intensity.

In conclusion it is indispensable to turn to one more parameter that obviously must be assigned to the same level in the processing of the signal. This parameter is the duration of the section (segment) of the signal.

The interest in this parameter consists in the fact that it makes obligatory a preliminary segmentation of the utterance into sections.

There is no doubt that a human listener somehow determines the duration of segments corresponding to vowels (it is possible to determine stress placement on the basis of comparing the durations of vowels in a sequence; in a number of languages the duration of vowels has phonemic significance). It is also known that the duration of the hold (approximation) of a consonant makes it possible to distinguish a double consonant from a single one.

There exist experimental data about the discrimination of the duration of the pause corresponding to the closure of a voiceless stop consonant [73, 74], and about imitating the duration of a voiced stop consonant in an isolated CV syllable [49]. On the basis of these data it is possible to assume that the value of the duration of a segment, established by the auditory system, is a monotonic function of its physical duration. In the phonetic interpretation of the obtained value supplementary information is employed, concerning probably the tempo of speaking and/or the duration of other nearby segments.

The question appears completely unexplored regarding the mechanism of auditory segmentation of the utterance. Therefore we can only list some assumptions bearing on this question.

The most obvious of these is the assumption that auditory segments need not coincide with phonemes in the sense that each segment contains information about one and only one phoneme and that the number of segments is equal to the number of phonemes.

One of the possible assumptions is that segmentation is accomplished as a result of processing the complete spectral-temporal description of the signal, and the points of segmentation are established at instances at which significant changes take place in the spectral picture. This would correspond, logically, to assuming that the isolation of segmentation signals proceeds at the same level as the isolation of parameters describing 'instantaneous' values of the spectral envelope and fine temporal structure. The functional meaning of segmentation signals might be that they control the consideration (transmission into short-term memory) of output signals from feature detectors, characterizing the dynamics of 'instantaneous' parameters during the extent of the segment.

Another proposition is that segmentation signals are formed as a result of processing 'instantaneous' parameters, and a separate segmentation might be performed for each of the parameters.

It is probable that the segmentation signals cannot be processed at a rapid rate. Thus, according to psycho-acoustic data, the temporal threshold of the perception of sequential ordering is approximately 20 msec [75].

5. The Phonetic Interpretation of Speech Stimuli

5.1. Units emerging from the block of phonetic interpretation.

Under the phonetic interpretation of a stimulus we understand the process of working out an auditory description of the stimulus, as a result of which a definite articulatory reaction may be associated with the stimulus. If large numbers of such reactions R_1 and R_2, observed in response to numerous repetitions of stimuli X_1 and X_2, do not differ among themselves, we accept that one and the same phonetic description, and one and the same phonetic image corresponds to both stimuli.

The phonetic image may be specified either as the set of instructions for synthesizing the speech complex in case we consider it from the point of view of the final stages of transformation in imitation, or as the designation of a multitude of stimuli (and a multitude of auditory descriptions) possessing certain given properties, in case we consider it from the point of view of initial stages of transformation [75].

Inasmuch as the phonetic image is an abstract description of both the acoustic stimulus and the motor complex, its internal structure must reflect the constraints that are essential both to the auditory system and to the system of speech production.

At the basis of contemporary linguistic investigations of language lies the assumption that the speech signal is described in perception and production in terms of a set of segmental units--phonemes, and suprasegmental units--prosodemes. This assumption is supported by a series of experimental data. Thus, a study of the mimicry of vowels [76] showed that in response to a signal, the subject selects one of a limited set of known configurations of the vocal tract. Thereby a certain category (multitude) of speech stimuli corresponds to each configuration, so that information about the required configuration may be represented in the phonetic image in the form of a symbol.

The contradiction, well known to engineers, between the linguistic approach to a phonemic system and the 'technical' (from the point of view of automatic speech recognition) description of phonetic images consists in the fact that the set of phonemes must be minimally small for a linguist, while the set of phonetic images need not meet this condition for an engineer. The requirement for economy may be left unsatisfied if it counteracts the requirement for reliability in recognition.

Cases in which the sets of phonemes and phonetic images do not coincide are found in instances in which one and the same phoneme is

realized in essentially different ways as a result of the influence of immediate phonetic context. As the most characteristic example of this we may consider Russian vowels after hard and soft consonants: ta - t'a, to - t'o, etc. [77, 78].

From the point of view of the reliability of automatic recognition, it is useful to describe separately the groups of vowels after hard and soft consonants and to assign to them separate symbols. An experimental study of the perception of these vowels showed that listeners (native speakers of Russian) proceed in exactly this manner--they interpret the a of ta and the a of t'a as separate entities, although from the linguistic point of view they constitute one phoneme /a/ [78].

At the present time enough data have been accumulated [83, 84] to maintain that the number of different entities used by the brain of a native speaker of Russian in the interpretation of vowels is larger than the number of vowel phonemes in the Russian language established at the linguistic level.

In order to designate these entities one might want to introduce some kind of new terminology, since they do not coincide completely with phonemes determined linguistically. The essential difference between them is that basically, from a linguistic point of view, redundant features of phonemes (those that arise, for example, as a result of the influence of some other phoneme) are not considered important for their description and isolation as separate phonemes, whereas for a listener it is indeed the redundant features that are made use of. In the following discussion the term 'phoneme' will be used, but the term will be understood to refer to the subjective image employed by the brain of the listener in the process of speech recognition. Other investigators use the term 'sound type' in this sense.

It has been shown experimentally that a number of subjective images--phonemes--really exists in the human nervous system, and that this number is not only finite but quite limited in size. Final data about the size of this number for native speakers of each concrete language are not yet available. As was mentioned above, only the minimal set of phonemes is established linguistically; it is unclear, however, whether any arbitrary linguistic phoneme can be associated with a phonetic image. For solving this problem it is indispensable to turn to methods of experimental psychology that have been worked out in the last few years (the method of mimicry, the analysis of confusion matrices, the method of active search for boundaries along phonetic categories, and methods of psychological scaling).

Let us consider how one might describe a phoneme taking it as a symbol denoting a certain range of auditory images and a certain articulatory complex. According to one of the methods for producing such a description, each one of the symbols is independent of every other symbol. According to another approach, the set of phonemes is systematically arranged, and each phoneme is described by listing the values of some of its 'distinctive' features. In this case, the number of features is significantly smaller than the number of phonemes. The idea of such a description of phonemes belongs to N. S. Trubetzkoy

[84]; it has been developed by R. Jakobson, M. Halle and G. Fant [31].

The logic of such propositions is rather obvious, if one looks at the phoneme as a set of instructions for the synthesis of an articulatory complex. These instructions must relate to speech organs (groups of muscles), and they can be considered as a set of elementary commands addressed to the various organs (vocal folds, lips, different muscle groups of the tongue etc.). One of the basic tasks of present-day physiological phonetics is to determine which sets of elementary instructions (motor commands) correspond to phonemes [85].

The idea, formulated by linguists, that phoneme sets are inherently systematic, also finds confirmation in specifically linguistic regularities (positional and conditioned sound changes, historical sequences of changes, morphological regularities, etc.). The description of such regularities appears more economical, if a phoneme is represented not as an isolated symbol, but as a list of its distinctive feature values.

Experimental proof that human listeners recall phonemes on the basis of a set of feature values was produced in investigations by Wickelgren [86, 87] and Galunov [82]. In these experiments listeners were presented series of six sound sequences, e.g. of the type CVC, where the vowels were different, but the consonants remained the same. The subjects had to write down the sequences after having heard the whole series. Mistakes made in writing down the recalled sequences were analyzed. It turned out that the mistakes have a very regular character. For each transmitted phoneme there exist some 'close' phonemes with which it is most frequently confused. This could not be the case, if the phonemes were remembered as isolated symbols, unconnected with any other symbols--phonemes.

Thus it is advantageous to accept that phonemic information, as it emerges from the block of phonetic interpretation, must be represented in terms of abstract distinctive features. Which must be the concrete set of these features and how many gradations are possible for each feature remains as yet unclear.

Very important is the question concerning the mutual connections between acoustic (auditory) features of the speech signal and the distinctive features of phonemes. The simplest and most attractive, although as yet experimentally unproven, is the proposition that distinctive features are binary, and that for each distinctive feature there exists a corresponding decision rule, its proper decisive boundary in the space of auditory features. If the auditory image that is called forth by the stimulus is located to one side of the boundary, the value of the stimulus according to the distinctive feature has one sign, and if it occurs on the other side of the boundary, its value according to the distinctive feature has the opposite sign.

It has been established sufficiently firmly at the present time that information concerning one and the same distinctive feature is contained in several acoustic (auditory) features of the stimulus [63, 89, 9]. This means that the decisive boundary may constitute a hypersurface in the space of these several auditory features. If the auditory features themselves are binary (cf. the preceding section),

the decisive boundary may have the maximally simple form

$$\sum_{1}^{i=n} k_i x_1 = 0,$$

where x_i - the auditory feature (+I or -I), k_i - the weight of the given auditory feature.

Recently obtained data concerning the decision rules that are employed in distinguishing between synthetic b and m [75] fit into this kind of a primitive scheme. It was also discovered that a human subject is able to give a numerical estimate of the closeness of the synthetic stimulus to the phoneme [75]. This makes it possible to admit that information concerning the distinctive feature at the output of the phonetic interpretation block determines not only the sign of the function

$$\sum_{1}^{i=n} k_i x_1,$$

but also its modulus. This is equivalent to saying that what is remembered is not a categorical decision concerning the class of phonemes (e.g. nasal or non-nasal) to which a given stimulus must belong, but the probability with which the stimulus may belong to this class.

The advantages of preserving this kind of information have already been discussed above (section 2.2.). It was experimentally proven by Lindner [90] that a final phonemic decision concerning an uncertain vowel may be made after the second vowel following the first one in time has been perceived.

5.2. The procedure of phonetic interpretation of auditory descriptions.

Most complex appears to be the question concerning the temporal organization of the process of phonetic interpretation. Direct experimental data concerning this question do not yet exist; however, some important requirements are known which must be met. One of them is that the procedure must ensure the collection of information that is contained in auditory features of different nature, distributed in time within the limits of approximately one syllable (cf. the surveys in [14, 90]). A second and most important requirement is that the failure to recognize an element must not lead to its being omitted--it must be indicated in the completed sequence of phonetic images that at a given point within the sequence there was an unrecognized (partially recognized) element [90].

The first of these requirements presupposes the existence of memory. Information regarding distinctive features (let this be the meaning of the function $\sum k_i x_i$) must accumulate with each occurrence

(in the general case, non-simultaneous occurrence) of the auditory features x_1, x_2, ... x_n. The second requirement presupposes an obligatory segmentation and breaking-up of the data. In the contrary case, information about the first, not yet completely recognized element will be mixed with information about the second element in the temporal sequence. As a result the first element will be left out, and the second may be incorrectly determined. It seems to us that one of the most important tasks for the immediate future consists of working out several different models for procedures that would satisfy both above-mentioned requirements, and of finding methods for their experimental verification.

Below we will attempt to describe--as yet in a very tentative manner--a hypothesis that appears plausible for a series of reasons, and, according to our opinion, requires further elaboration and testing. It might be called the hypothesis of syllable recognition.

We propose that the process of phonetic interpretation includes the operation of a special program that marks off syllables (open syllables). It is clear that the isolation of elements constituting rhythmic and melodic structures takes place during the perception of very varied and not even necessarily speech-like signals. The rhythmic and melodic structure of a phrase may be transmitted by means of signals that are very remote from speech sounds. There exist data that some patients with sensory aphasia have no difficulty in reproducing the rhythmic structure of a phrase, although they can neither understand it nor reproduce the sequence of phonetic elements of which it is constituted [6].

Under very high spectral distortion and correspondingly very low phonemic intelligibility of speech, the perception of its rhythmic structure is almost unaffected [90]. This makes it very plausible to assume that in speech perception two independent procedures are employed in parallel. One is responsible for the segmentation of the stream of speech into syllables (elements of rhythmic sequence) and the description of the so-called prosodic characteristics of the sequence, the other is responsible for the description of the characteristics of each separate syllable, which is accomplished in terms of phonemes or distinctive features.

We just used the term 'segmentation' of the stream of speech. Since it is frequently used in very different meanings, it is indispensable to dwell somewhat more specifically on what we have in mind. As of now we propose only that as a result of some kind of a procedure every syllable is associated with a kind of 'mark' (impulse), so that the number of impulses that arise in the process of listening to a sentence will be equal to the number of syllables in that sentence.

From the fact that a human is able to repeat a meaningless sequence of 7 - 10 syllables without confusing their order in time and without distorting the prosody, it follows that when perceived information is registered in memory certain reference signals must be employed that allow one to group together phonetic and prosodic information about the syllable and assign the syllable its order number. The role of such reference signals must be played by the

proposed syllable impulses. It is possible, for example, to assume that the short-term memory into which the speech sequence is entered consists of K cells which are filled in sequence. The syllable impulse performs the role of a switching signal, switching the output of the preceding level of the system from one cell to the next.

Allowing for obligatory switching makes it possible not to make final phonemic decisions if sufficient information is not available during the extent of the syllable; it makes the possibility of prolonged preservation of information regarding the input stimulus compatible with the absence of confusion with data relating to successive phonemes.

Up to now we have said that a separate memory cell corresponds to each successive syllable. However, we have also said that information within the cell is entered in terms of phonemes. Output signals of the preceding level correspond to a running description of the stimulus in terms of auditory features. In order to proceed from one kind of description to the other it is necessary to use some kind of a decoder.

Our next proposition is that within the nervous system there exists a series of identically organized decoders (their number is equal to the number of syllables that can be kept in memory simultaneously). The syllable impulse accomplishes the successive switching of input information from one decoder to the other. Each separate decoder accomplishes the transition from the sequence of auditory features present during an open syllable to the description of this open syllable in terms of phonemes or distinctive features.

The fact that the decoder must be designed to operate on open syllables follows, firstly, from the fact that speech is articulated as a sequence of open syllables (the articulation of the vowels begins simultaneously with the articulation of the consonant [91, 92]) and, secondly, from the observation that the interpretation of the stimulus during the consonant part depends on the properties of the stimulus during the part of the following rather than the preceding vowel [14, 93].

The use of open syllables in the capacity of input signals to the decoder appears very reasonable both from the point of view of the procedure of phonemic recognition and from the point of view of the relatively low requirements to be made in this case with regard to the procedure of segmentation. The collection of information about the phoneme may be performed during the whole temporal segment in which this information is actually present; stationary and transitional parts may be utilized equally. The number of elements in the output alphabet of the decoder may be approximately equal to the number of phonemes, since the contextual mutual influences may be accounted for in the decision rules themselves.

According to the motor theory of perception, the work of the decoder consists in transforming the perceived signal into a set of motor commands which would be required for imitating what is heard. How is the selection of required motor commands carried through? It is difficult to suppose that this is done by the method of

surveying the complete set of hundreds of thousands of possible variants of syllables. A parallel survey would demand a great expenditure of functional elements, and a sequential one--a great amount of time.

It is possible to decrease the number of variants on the basis of a preliminary recognition of units which are simply connected with elements of motor complexes. These units, obviously, must be close to phonemes. It may turn out that the reliability of recognizing phonemes on the basis of their acoustic characteristics will not be high. It is possible to imagine that in this case certain of the most probable phonemes will stand out with the indication of their probabilities. After this, motor complexes are formed which are required for imitating sequences of these most probable phonemes. The number of possible variants of such sequences will be smaller by several orders of magnitude than the initial number of possible variants. From these variants the variant will be selected that possesses the maximal production values for three quantities: the a posteriori probability of phonemes, their a priori probability, and the a priori probability of the sound sequence. The last two values reflect knowledge of the laws of the language. In this manner decisions about the recognition of sound types are made more precise simultaneously with the recoding of the acoustic signal into motor commands, i.e., into a very compact code.

The procedure described here may explain one of the peculiarities of motor perception of speech. At the same time it is a short description of the above-mentioned algorithm for increasing the reliability of recognition on the basis of the redundancy of the signal [21].

Let us return to segmentation. The basic requirement to be set up for this procedure is that information pertaining to a phoneme in one syllable must not be attributed to a phoneme of another syllable. It is granted that the syllable impulse arises somewhere in the transition from vowel to consonant. Omission of the transitional section is not dangerous from the point of view of recognizing the consonants of the second syllable, since the transition contains very little information [79, 83]; the possibility of attributing this section to the consonant of the preceding syllable can be easily excluded on the basis of limitations incorporated in the schema of the decoder itself (the consonant following the vowel is excluded).

It seems a priori obvious that the most complicated task in working out a model of this type is the recognition of consonant clusters. From this point of view it appears extremely important to obtain experimental data regarding the perception of consonant clusters in nonsense sound sequences.

6. Scheme of a Model of Speech Perception

Basing ourselves on all facts and assumptions discussed above, let us now attempt to outline the most plausible general scheme of a mechanism intended for the recognition of a sufficiently large

quantity (2 - 3 thousand) of spoken words. A speech signal, constituting a non-stationary function of time f(t), arrives at the input of the mechanism. The output block must issue a decision about the assignment of the unknown realization to one of the earlier indicated 2 - 3 thousand words of the lexicon S_o with a reliability P_o, which is comparable to the reliability with which these speech signals are perceived by human listeners. It is clear that this mechanism must have a hierarchical structure of the type represented on Fig. 1. It is indispensable to make more precise the number of stages (elementary automata), make more concrete the contents of each stage and describe the procedure for processing the signal in its progress from the input to the output of the mechanism.

It follows from the foregoing that at each hierarchical level a block may be isolated that carries through the perception procedure (receptor X_i), a decision-making block (classifier D_i) and a block stating the decisions made with the reliability P_i (effector S_i).

It may be expected that because of the limited abilities of the classifiers D_i the recognition of elements S_i will be performed with an unacceptably low reliability P_i. It would be useful to have at every level blocks (H_i) for the correction of errors. Errors may be eliminated on the basis of a priori information about speech and language, which may be stored in long-term memory. This information is of different kinds--it may constitute knowledge about limitations in the physical characteristics of the speech production apparatus or about linguistic regularities in the language.

Taking into account what has been said, the procedure for recognizing elements at one of the levels may look like the following [95].

The classifier D_i indicates some hypotheses S_i to which the vector of unknown realization X_i may be attributed with the greatest probability. It is logically inevitable that a certain block Q_i (let us call it 'supervisor') be present, which must evaluate the quality of decisions being made and, according to necessity, include reserves of one kind or another for increasing the reliability of recognition. The evaluation of the quality may consist in the simplest case of the determination of the difference ΔP_i of the a posteriori probabilities of competing hypotheses S_i. A decision is considered satisfactory, if ΔP_i exceeds a certain fixed threshold V.

Increasing the reliability may be achieved by providing a more and more complete description of the realization X_i, i.e., by the analysis of a wider range of parameters. After that, when these possibilities have been exhausted, and $\Delta P_i < V$, the supervisor includes block H_i (the error-correction block) on the basis of a priori information about the characteristics of the speech tract or linguistic regularities.

As Wald has shown [96], this sequential procedure for increasing the reliability of recognition ensures minimal mathematical expectation of cost for making the decision.

It should be mentioned that the order in which these or other means for increasing the reliability of recognition are adopted depends on the relationship between the useful result of a given method

and the cost of its realization. This relationship is as yet unknown to us; therefore the set and the order of inclusion of means by the supervisor may differ from what has been described above.

If at any step the probability of a hypothesis exceeds the probability of any other hypothesis by more than V, then this hypothesis S_i is transmitted to the input of the following (i + 1th) stage of recognition. In the contrary case several of the most probable hypotheses are retained in memory, which are then transmitted to the input of the i + 1th level one after the other, in the order of decreasing probability. It is possible to imagine another--parallel--method, according to which all competing hypotheses enter simultaneously at the inputs of several classifiers of the same type at the i + 1th level.

If in the sequential scheme economy is achieved with respect to the number of functional elements, then in the parallel scheme the time needed for decision-making is decreased. The effectiveness of parallel application of algorithms for solving complex tasks in computing systems [97] points to the usefulness of a parallel scheme for processing information, especially when it is necessary to obtain high productivity on the basis of slowly acting functional elements; however, we do not yet know direct experimental facts in favor of one or the other scheme for the processing of information by the brain.

Facts presented in the beginning of this work speak in favor of the assumption that on the level at which phonemes are recognized, a decision is made taking into account information scattered over a segment of the type of an open syllable. It follows from here that in the scheme of the automaton there must be a block for the segmentation (C_i) of the stream of speech into open syllables. It is probable that blocks devoted to the segmentation of the speech stream into one or another kind of sections must be present also at other hierarchical levels. Thus besides X, D, S, P there must be present blocks Q, H, and C. What will the procedure of processing the speech signal now look like, when it passes through the recognition mechanism? In the sequential variant (Fig. 3) the speech signal f(t) is transformed at the very beginning into a rather complete description in the space (X) of frequency and time. During a section ('window') of a certain duration, determined by the short-term memory capacity of the input chains of the auditory analyzer, some features (S_1) are isolated of the type of static and dynamic characteristics of formants, characteristics of the noise part of the spectrum etc., normalized for loudness, tempo and some other parameters. It is possible that this procedure is articulated into a series of smaller stages, as, for example, loudness normalization, isolation of static characteristics, tempo normalization, determination of dynamic characteristics, etc.

In technical models the segmentator C_1 may be needed for establishing the boundaries of the temporal 'window'.

The indispensable reliability of recognition (P_1) of features may be obtained by using information about physical laws of speech production of the following type: the frequency of the fundamental cannot be changed faster than at a certain speed; simultaneous existence of such and such features is impossible: after a given

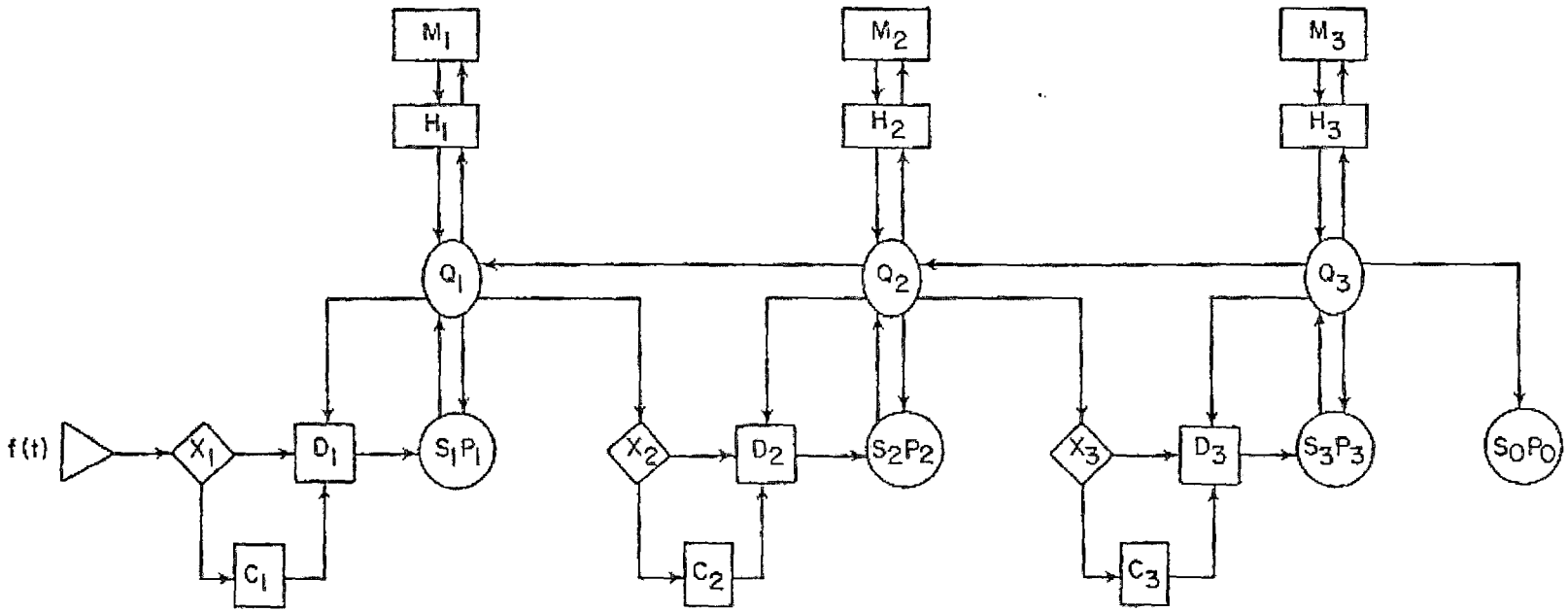
M1
M2
M3
H1
H2
H3
Q1
Q2
Q3
f(t)
X1
D1
S1P1
X2
D2
S2P2
X3
D3
S3P3
S0P0
C1
C2
C3

Fig. 3

combination of features, the occurrence of another given set of features is most probable, etc.

In living systems this information is probably included in the construction of blocks that measure the characteristics of a speech signal in the form of time constants, bandwidths, schemes for suppressing or sharpening various maxima etc. In technical mechanisms, information about speech production may be stored in the long-term memory M_1.

The use of this kind of information is continued until the probability of a certain variant of features becomes greater than the probability of other variants relative to a certain threshold value ∇. This variant enters at the input of the next stage of transformation, where the sequence of such features constitutes the space of description X_2.

If the difference between probabilities is smaller than ∇, then several variants of features S_1 are kept in memory.

On the second level the recognition of phonemes (S_2) takes place. For this purpose, the classifier D_2 employs information from the open syllable type segment, whose boundaries are determined by the segmentator C_2. For reducing the number of possible variants of a certain phoneme, information is used that is contained in the description X_2, and afterwards, if necessary, also information from M_2 concerning the structure of phoneme sequences. For this purpose, block H_2 formulates sequences of most probable variants of phonemes and, taking into consideration all this a posteriori and a priori knowledge, selects the most probable sequence. If the difference in probability of this selected sequence and any arbitrary sequence exceeds a certain threshold ∇, then the phonemic code of the syllable is transmitted to the input of the following block. In the contrary case, a categorical decision is not made and the phoneme codes S_2 of several (most probable) syllables are retained in memory. If there are too many variants, the procedure may be repeated, calling forth another variant of features X_2 along the line $Q_2 - Q_1$ at the input of the block.

In order to recognize words from the lexicon S_0 the space X_3 must contain, in addition to phoneme codes, information about stresses. The segmentator C_3 carries through the segmentation of the speech stream into sections stretching from one stress to the next. Two such neighboring sections contain as a minimum one word of the lexicon S_0. The search for the needed word and the simultaneous determination of its boundaries may be accomplished by means of the algorithm of Lisenko [24]. At this stage as well as earlier, additional a priori knowledge from M_3 about the elements of the lexicon may be used (block H_3) in the selection of a decision, and if it should prove indispensable, other variants of the phonemic sequences may be summoned (along the line $Q_3 - Q_2$) to the input (X_3).

Differing from this, in the scheme with blocks working in a parallel mode (Fig. 4) several most probable variants of features S_1 are simultaneously transmitted to the input of the second level. In each of α branches the classifier D_2 (j) establishes whether the vector $X_2^{(j)}$ belongs to one of the phonemes of the alphabet S_2.

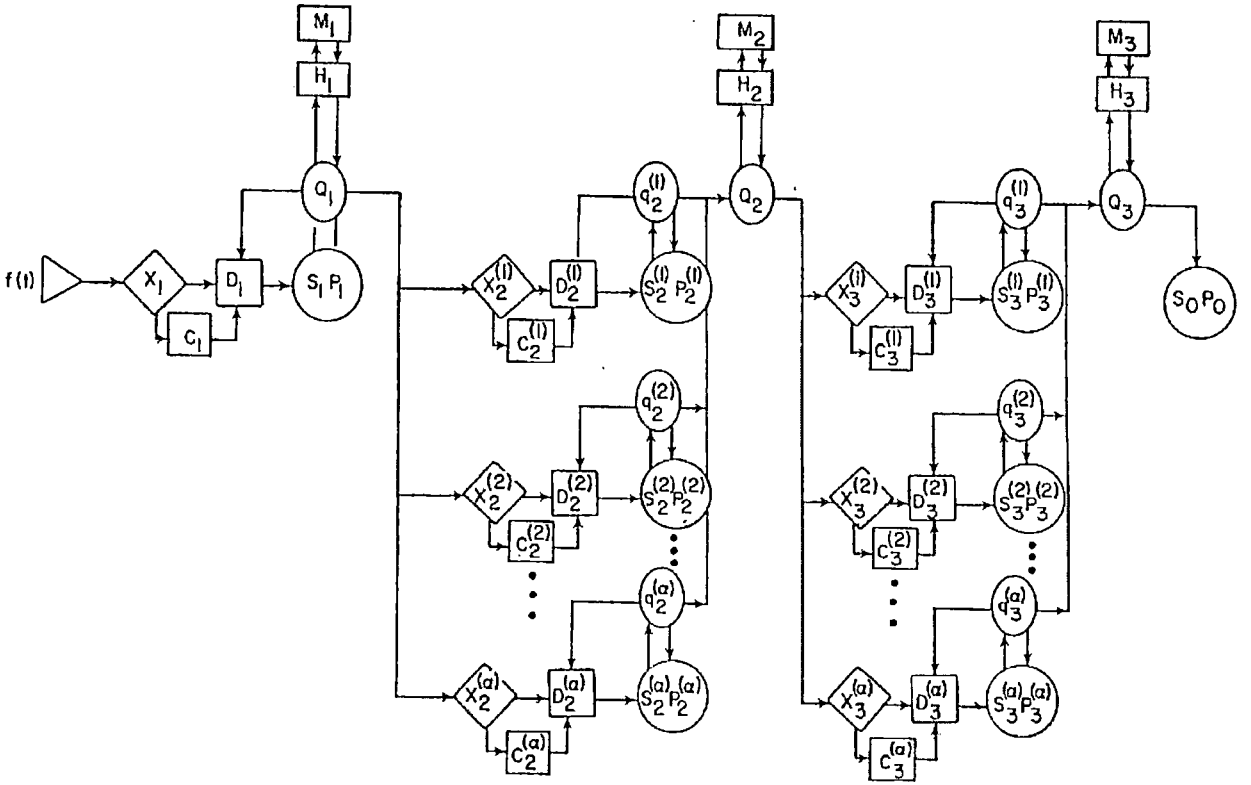
f(1)
X1
C1
D1
S1 P1
Q1
M1
H1
X2(1)
C2(1)
D2(1)
S2(1) P2(1)
q2(1)
X2(2)
C2(2)
D2(2)
S2(2) P2(2)
q2(2)
X2(a)
C2(a)
D2(a)
S2(a) P2(a)
q2(a)
Q2
M2
H2
X3(1)
C3(1)
D3(1)
S3(1) P3(1)
q3(1)
X3(2)
C3(2)
D3(2)
S3(2) P3(2)
q3(2)
X3(a)
C3(a)
D3(a)
S3(a) P3(a)
q3(a)
Q3
M3
H3
S0 P0

Fig. 4

The most probable hypotheses are transmitted by the supervisors $q_2^{(j)}$ to the input of supervisor Q_2, which functions in the same way as Q_2 in the sequential variant.

The peculiarities of the functioning of blocks working in parallel at the third level of recognition are analogous.

These schemes do not contradict presently known facts about human speech perception. At the same time we are conscious of the possibility that a further development of investigations in this area may lead either to a concretization of these schemes or to a necessity for changing them in very basic ways.

We believe that the tasks immediately ahead consist of further investigations of the structure, methods of functioning and interactions of human prototypes of the blocks which enter the schemes presented above.

BIBLIOGRAPHY

1. Gud G.X., Makol R.E. Sistemotexnika. Vvedenie v proektirovanie bol'šix sistem. Perevod s angl. Izd-vo "Sov. radio". M., 1962.

2. Mel'čuk I.A. Avtomatičeskij sintaksičeskij analiz. Novosibirsk, 1964.

3. Lejkina B.N., Nikitina T.N., Otkupščikova M.I., Fitialov S.Ja., Cejtlin G.S. Sistema avtomatičeskogo perevoda, razrabatyvaemaja v gruppe matematičeskoj lingvistiki. V.C. LGU "NTI", 1966, I, 40-50.

4. Simmons R.F. ECVM, otvečajuščaja na voprosy, zadannye no-anglijski. Zarubežnaja radioelektronika, 1965, 7. 49-82.

5. Galunov V.I., Čistovič L.A.O svjazi motornoj teorii s obščej problemoj raspoznavanija reči. Akustičeskij žurnal, 1965, II, 417-426.

6. Stevens K.N. and House A.S. Speech perception in foundations of modern auditory theories (eds. S. Tobias a.E. Schubert), in press.

7. Fant G. Auditory patterns of speech. Proc. of the Symposium on Models for the Perception of Speech and Visual Form. Boston, Mass., Nov. 11-14, 1964.

8. Chistovich L.A., Golusina A., Lublinskaja V., T. Malinnikova, M. Zhukova. Psychological methods in the speech perception research. Zeitschrift für Phonetik, 1968, in press.

9. Fant G. The nature of distinctive features. Speech Trans. Lab. Quart. Progr. a. Status Rep., 1966, 4, 1-14.

10. Nil'son N. Obučajuščiesja mašiny. Izd-vo "Mir", M., 1967.

11. Miller G.A. The magical number seven, plus or minus two: some limits in our capacity for processing information. Psychol. Rev., 1956, 63, 81-97.

12. Nevel'skij P.B. Sravnitel'noe issledovanie ob'ema kratkovremennoj i dolgovremennoj pamjati. 18 meždunarodyj psixologičeskij kongress. Simpozium 21, 1966, 26.

13. Čistovič L.A., Klaas Ju.A., Alëkin R.O. O značenii imitacii dlja raspoznavanija zvukovyx posledovatel'nostej. Vopr. psixologii, 1961, 5, 173-182.

14. Čistovič L.A., Koževnikov V.A., Aljakrinskij V.V. i dr. Reč'. Artikuljacija i vosprijatie. "Nauka", 1965.

15. Mehler S. Some effects of grammatical transformations on the recall of English sentences. S.verb. Learn.verb. behaviour, 1963, 2, 346-351.

16. Zagorujko N.G. Kombinirovannyj metod prinjatija rešenij. Sb. tr. IM SO AN SSSR. "Vyčislitel'nye sistemy", vyp. 22, Novosibirsk, 1966.

17. Zagorujko N.G. Kakimi rešajuščimi funkcijami pol'zuetsja čelovek? Sb. tr. IM SO AN SSSR. "Vyčislitel'nye sistemy", vyp. 28, Novosibirsk, 1967.

18. Čistovič L.A., Fant G., Serpa-Lejta A., T'ernlund P. Kvartal'nyj otčet laboratorii peredači reči Stokgol'mskogo Korolevskogo texnologičeskogo instituta, No. 4, 1966.

19. Miller G.A. Decision units in the perception of speech. IRE Trans. on Information Theory, 1962, 8, 81-83.

20. Galunov V.I. Nekotorye osobennosti vosprijatija reči. Akust. ž. 1966, 12, 422-427.

21. Vološin G.Ja. Ob ispol'zovanii jazykovoj izbytočnosti dlja povyšenija nadežnosti avtomatičeskogo raspoznavanija rečevyx signalov. Sb. tr. IM SO AN SSSR "Vyčislitel'nye sistemy", vyp. 28, Novosibirsk, 1967.

22. Zagorujko N.G., Lozovskij V.S. Podstrojka pod diktora pri raspoznavanii ograničennogo nabora ustnyx komand. Sb. tr. IM SO AN SSSR, "Vyčislitel'nye sistemy", vyp. 28, Novosibirsk, 1967.

23. Fujisaki H. a. T. Kawashima. The roles of pitch and higher formants in the perception of vowels. 1967 Conference on Speech Communication and Processing, 251-256.

24. Lisenko D.M. Principy vydelenija i morfologičeskogo analiza slova pri vosprijatii ustnoj reči. Diss., L., 1966.

25. Flanagan J.L. Speech, Analysis, Synthesis and Perception. Academic Press, New York, 1965.

26. Molčanov A.P., Labutin V.K. Slux i analiz signalov. Izd. Energija, 1967.

27. Glezer V.D. Mexanizmy opoznanija zritel'nyx obrazov. Nauka, 1966.

28. Al'tman Ja.A., I.A. Vartonjan, G.V. Geršuni, A.M. Maruseva, E.A. Radionova, G.I. Ratnikova. O funkcional'noj klassifikacii po vremennym xarakteristikam nejronov stvolovyx otdelov sluxovoj sistemy. Doklad, pročitannyj na Naučnoj sessii instituta fiziologii im. I.P. Pavlova AN SSSR, posvjaščennoj 50-letiju Velikoj Oktjabr'skoj Socialističeskoj revoljucii. Oktjabr' 1967 g.

29. Hubel D.H. and T.N. Wiesel. Receptive fields of single neurons in the cat's striate cortex. Journ. Physiol., 148, 574.

30. Whitfield J.C. and Evans E.F. Responses of auditory cortical neurons to stimuli of changing frequency. J. Neurophysiology, 1965, vol. 28, 655-672.

31. Jakobson R., Fant G., Halle M. Preliminaries to speech analysis. The distinctive features and their correlates. Acoust. Lab. M.I.T. Techn. Rep. 13. Cambridge, Mass., 1952.

32. Bondarko L.V. Differencial'nye priznaki slogov. Otčet lab. eksperimental'noj fonetiki. L., 24, 1967.

33. Bondarko L.V. Struktura sloga i xarakteristiki fonem. Voprosy jazykoznanija, No. 1, 1967.

34. Varšavskij L.A. Ob avtomatičeskom raspoznavanii reči. Voprosy radioelektroniki. ser. XI, vyp. 1, 1964, 5-22.

35. Paul A.P., House A.S. and Stevens K.N. Automatic reduction of vowel spectra: an analysis-by-synthesis method and its evaluation. J.Acoust.Soc.Am., 1964, 36, 303-308.

36. Čistovič L.A. Psixoakustika i voprosy teorii vosprijatija reči. Raspoznavanie sluxovyx obrazov. "Nauka", Novosibirsk, 1966, 68-168.

37. Šupljakov V.C. O tonal'noj vysote zvukov /s/ i /š/. Sb. "Mexanizmy rečeobrazovanija i vosprijatija složnyx zvukov." 1966, 87-95.

38. Šupljakov V.C. Sluxovoj analiz stacionarnyx šumnyx soglasnyx. Diss. L., 1967.

39. Šupljakov V.C. Akustičeskij priznak vosprijatija mjagkosti stacionarnyx šumnyx soglasnyx. VI Vsesojuznaja akustičeskaja konferencija. M., 1968.

40. Chistovich L.A., Fant G. A. de Serpa-Leitao. Mimicring and perception of synthetic vowels. Part II. Speech Transm. Lab. Quart. Progr. a. Status Report. Stockholm, 1966, 3, 1-3.

41. Small A.M. and Daniloff R.G. Pitch of noise bands. J. Acoust. Soc. Am., 1967, 41, 506-512.

42. Ekdahl A.G. and Boring E.G. The pitch of tonal masses. Am. J. Psychol., 1934, 46, 452-455.

43. Nábělek I., Krútel J., Majerník V. Ein Beitrag zur Bestimmung der Tonhöhe von Bandpassrauschen. III Akust. Konferencia. Budapest. 308-313.

44. Plomp R. The ear as a frequency analyzer. J. Acoust. Soc. Am., 1964, 36, 1628-1638.

45. Flanagan J.L. A difference limen for vowel formant frequency. J. Acoust. Soc. Am., 1955, 27, 613-617.

46. Miller G. A. and Nicely P.E. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am., 1959, 27, 338-352.

47. Čistovič L.A. Vlijanie častotnyx ograničenij na razborčivost' russkix soglasnyx zvukov. Sb. Telefonnaja akustika. L., 1955, 1-2, 35-42.

48. Derkač M.F. Statistika vosprijatija gluxix vzryvnyx i ščelevyx soglasnyx v zavisimosti ot ix dlitel'nosti. Sb. "Voprosy statistiki reči". L., 1958, 40-45.

49. Žukova M.G. Vosprijatie i vosproizvedenie dlitel'nosti sintetičeskogo soglasnogo v sloge tipa SG. 1968. Trudy VI Vsesojuznoj akustičeskoj konferencii.

50. Fant G. Akustičeskaja teorija rečeobrazovanija. "Nauka", 1964.

51. Čistovič L.A. Psixofizičeskie xarakteristiki sluxa. "Inženernaja psixologija". Izd. MGU, 138-158.

52. Flanagan J.L. Audibility of periodic pulses and a model for the threshold. J. Acoust. Soc. Am., 1961, 33, 1540-1549.

53. Tumarkina L.N., N.A. Dubrovskij. Nekotorye osobennosti vosprijatija čelovekom amplitudno-modulirovannyx signalov. Biofizika, 1966, II, v.4, 653-658.

54. Borovičová B. and Maláč V. Towards the basic units of speech from the perception point of view. Proc. Seminar on Speech Production and Perception. Leningrad, 1966. Z. für Phonetik usw. in press.

55. Ohman S.E.G. Numerical model of coarticulation. J. Acoust. Soc. Am., 1967, 41, 310-320.

56. MacNeilage P.F. and De Clerk J.L. On the motor control of coarticulation in CVC monosyllables. 1967, Conference on Speech Communication and Processing. Boston, Mass., C3, 157-163.

57. Sorokin V.N., Fajn V.S. Nepreryvno-gruppovoe raspoznavanie slov: algoritm i eksperimental'nye resul'taty. Trudy VI Vsesojuznoj akustičeskoj konferencii. 1968. M.

58. Stevens K.N., House A.S., Paul A.P. Acoustic description of syllabic nuclei: an interpretation in terms of a dynamic model of articulation. J. Acoust. Soc. Am., 1966, 40, 123-132.

60. Liberman A.M., Delattre P., Cooper F.S. and Gerstman L.J. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. Psychol. Monographs, 1954, 68, No. 8, p. 1-13.

61. Delattre P., Liberman A.M. and Cooper F.S. Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am., 1955, 27, 768-773.

62. Harris K.S., Hoffman H.S., Liberman A.M., Delattre P., Cooper F.S. Effect of third-formant transitions on the perception of the voiced stop consonants. J. Acoust. Soc. Am., 1958, 30, 122-126.

63. Hoffman H.S. Study of some cues in the perception of the voiced stop consonants. J. Acoust. Soc. Am., 1958, 30, 1035-1041.

64. Öhman S.E.G. Coarticulation in VCV utterances: Spectrographic measurements. J. Acoust. Soc. Am., 1966, 33, 151-168.

65. Stevens K.N. Acoustic correlates of certain consonantal features. 1967 Conference on Speech Communication and Processing. Boston, Mass., C6: 177-184.

66. Rejtblat L.E. Vosprijatie napravlenija izmenenija častoty spektral'nogo maksimuma v sintetičeskom sloge. Diplomnaja rabota, 1968.

67. Liberman A.M., Delattre P.C., Gerstman L.J. and Cooper F.S. Tempo of frequency change as a cue for distinguishing classes of speech sounds. J. exp. Psychol., 1956, 52, 127-137.

68. Brady P.T., House A.S. and Stevens K.N. Perception of sounds characterized by a rapidly changing resonant frequency. J. Acoust. Soc. Am., 1967, 33, 1357-1362.

69. Heinz J.M., Lindblom B.E.F., J.Ch.K-G. Lindquist. Patterns of residual masking for sounds with speech-like characteristics. 1967 Conference on Speech Communication and Processing. Boston, Mass., D1, 246-251.

70. Čistovič L.A. Izmenenie osnovnoj častoty golosa kak različitel'nyj priznak soglasnyx. Akus. Ž. v pečati.

71. Heinz J. and Stevens K.N. On the properties of voiceless fricative consonants. J. Acoust. Soc. Am., 1961, 33, 589-596.

72. Liberman A., Harris K.S., Eimas P., Lisker L., Bastian J. An effect of learning on speech perception: the discrimination of duration of silence with and without phonemic significance. Language and Speech, 1961, 4, 175-195.

73. Huggins A.W.F. How accurately must a speaker time his articulation 1967, Conference on Speech communication and Processing. Boston, Mass., D6, 268-273.

74. Hirsh I.J. Auditory perception of temporal order. J. Accust. Soc. Am., 1959, 31, 753-767.

75. Čistovič L.A. O procedure raspoznavanija fonem čelovekom. Voprosy psixologii (v pečati).

76. Chistovich L., Fant G., A. de Serpa-Leitao, Tjernlund P. Mimicring of synthetic vowels. Speech Transmission Lab. Quarterly Progress and Status Report, 1967, 2, 1-18.

77. Bondarko L.V. O xaraktere izmenenija formantnogo sostava russkix glasnyx pod vlijaniem mjagkosti sosednix soglasnyx. Uč. zap. LGU, 1960, No. 237, vyp. 40.

78. Bondarko L.V. i Zinder L.R. O nekotoryx differencial'nyx priznakax russkix soglasnyx fonem. Voprosy jazykoznanija. No. 1, 1966.

79. Verbickaja L.A. Zvukovye edinicy russkoj reči i ix sootnošenie s ottenkami.i fonemami. Kand. diss., L., 1965.

80. Čistovič L.A. Klassifikacija zvukov reči pri ix bystrom povtorenii. Akust. ž., 1960, 6, 392-398.

81. Aljakrinskij V.V. Imitacija det'mi (4-7 let) russkix i nekotoryx anglijskix glasnyx. Voprosy psixologii, 1963, 4, 80-87.

82. Galunov V.I. Struktura množestva rečevyx obrazov. Diss. L., 1967.

83. Bondarko L.V., Verbickaja L.A., Zinder L.R. i Pavlova L.P. Različaemye zvukovye edinicy russkoj reči. Sb. "Mexanizmy rečeobrazovanija i vosprijatija složnyx zvukov." 1966.

84. N.S. Trubeckoj. Osnovy fonologii. M., 1960.

85. Liberman A.M., Cooper F.S., Studdert-Kennedy M., Harris K.S., D.P. Shankweiler. Some observations on the efficiency of speech sounds. Zeitschrift für Phonetik, 1968, in press.

86. Wickelgren W.A. Distinctive features and errors in short-term memory for English vowels. J. Acoust. Soc. Am., 1965, 38, 583-588.

87. Wickelgren, W.A. Distinctive features and errors in short-term memory for English consonants. J. Acoust. Soc. Am., 1966, 38, 388.

89. Delattre P. From acoustic cues to distinctive features. Rezjume dokladov 6-go Meždunarodnogo kongressa foneticeskix nauk. Praga 1967, 33.

90. Lindner G. Veränderung der Beurteilung synthetischer Vokale unter dem Einfluss des Sukzessivkontrastes. Zeitschrift für Phonetik, 1966, 19, N. 4/5 - 297-307.

91. Kok N.P. Izbiratel'noe rasstrojstvo v organizacii reči: nestojkost' sluxo-rečevyx sledov. Voprosy psixologii. 1965, 2.28-34.

92. Öhman S.E.G. Numerical model of coarticulation. J. Acoust. Soc. Am., 1967, 41 (2), 310-320.

93. Bondarko L.V., Zinder L.R., Pavlova L.P. Različaemye zvukovye tipy russkix soglasnyx. Voprosy radioelektroniki. TRS, vyp. 5, 1967.

94. Bondarko L.V. Vosprijatie differencial'nyx priznakov i slogovaja struktura reči. Tezisy dokladov VI Meždunarodnogo kongressa fonetičeskix nauk, Praga, 1967.

95. Zagorujko N.G. Problema raspoznavanija reči kak složnaja sistema. Tezisy dokladov VI Meždunarodnogo kongressa fonetičeskix nauk. Praga, 1967.

96. Wald A. Statistical decision functions. John Wiley & Sons, N.Y., 1950.

97. Evreinov E.V., Kosarev Ju. G. Odnorodnye universal'nye vyčislitel'nye sistemy vysokoj proizvoditel'nosti. Izd-vo "Nauka" Sib. otd. Novosibirsk, 1966.